REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
08-03-2008	Final Report	1-Jul-2003 - 30-Nov-2007

4. TITLE AND SUBTITLE
Heterogeneous Uncertainty Manageemnt

5a. CONTRACT NUMBER
DAAD19-03-1-0202

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102
611102

6. AUTHORS
V.S.Subrahmanian

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Maryland - College Park
Office of Research Administration & Advancement
University of Maryland, College Park
College Park, MD 20742 -5141

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
45443-CI.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; Federal purpose rights

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
We developed methods to study the uncertainty that arises when we need multiple forms of reasoning and when multiple data representations are involved. We developed the concepts of heterogeneous temporal probabilistic (HTP) agents, the concept of probabilistic version of XML and RDF, and probabilistic methods to reason about collections of moving objects.

15. SUBJECT TERMS
probabilistic logic, uncertainty, heterogeneous reasoning

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	SAR		V. Subrahmanian
					19b. TELEPHONE NUMBER
					301-405-6722

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Heterogeneous Uncertainty Manageemnt

ABSTRACT

We developed methods to study the uncertainty that arises when we need multiple forms of reasoning and when multiple data representations are involved. We developed the concepts of heterogeneous temporal probabilistic (HTP) agents, the concept of probabilistic version of XML and RDF, and probabilistic methods to reason about collections of moving objects.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Y. Zhang, E. Manister, S. Kraus, V.S. Subrahmanian, D. Peleg. Computing the survivability of a multi-agent deployment, Artificial Intelligence journal, accepted subject to revisions, Oct. 2007.

A. Parker, J. Grant, and V.S. Subrahmanian. A Logical Formulation of Probabilistic Spatial Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, Nr. 11, pages 1541-1556.

M. Fayzullin, V.S. Subrahmanian, M. Albanese, C. Cesarano, A. Picariello.
Story Creation from Heterogeneous Data Sources, Multimedia Tools and Applications, Vol. 33, Nr. 3, pages 351-377, June 2007.

E. Hung, L. Getoor and V.S. Subrahmanian. Probabilistic Interval XML, ACM Transactions on Computational Logic, Vol. 8, Nr. 4, August 2007.

J. Dix, S. Kraus and V.S. Subrahmanian. Heterogeneous Temporal Probabilistic Agents, ACM Transactions on Computational Logic, Vol. 7, Nr. 1, Jan. 2006.

R. Ross, V.S. Subrahmanian, and J. Grant. Aggregate Operators in Probabilistic Databases, Journal of the ACM, Vol. 52, Nr. 1, pages 54-101, January 2005.

J. Grant, G. Infantes, A. Parker and V.S. Subrahmanian. SPOT Databases: Efficient Consistency Checking and Optimistic Selection in Probabilistic Spatial Databases, submitted to a technical journal, Nov. 2007. Under revision in accordance with reviewers' comments.

Number of Papers published in peer-reviewed journals: 7.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

Peer-Reviewed Conference Proceeding publications (other than abstracts):

V. Martinez, A. Pugliese, G. I. Simari, V.S. Subrahmanian and H. Prade. How Dirty is your Relational Database?, Proc. ECSQARU 2007, pages 103-114.

O. Udrea, Z. Majkic and V.S. Subrahmanian. Aggregates in generalized temporally indeterminate databases, Proc. 2007 International Conference on Scalable Uncertainty Management, Lecture Notes in Computer Science Vol. 4772, pages 171-186, Springer Verlag.

M. Albanese, A. Pugliese, V.S. Subrahmanian, and O. Udrea. MAGIC: A Multi-Activity Graph Index for Activity Detection, Proc. 2007 IEEE International Conference on Information Reuse and Integration, Las Vegas, August 2007. pages 267-272.

M. Albanese, V. Moscato, A. Picariello, V.S. Subrahmanian, O. Udrea. Detecting Stochastically Defined Activities in Video, Proc. IJCAI 2007, pages 1802—1807, Jan. 2007, Hyderabad, India.

F. Yaman, A. Parker and V.S. Subrahmanian. Probabilistic Go-Theories, Proc. IJCAI 2007, pages 501-506, Jan. 2007, Hyderabad, India.

V.S. Subrahmanian and L. Amgoud. A General Framework for Reasoning about Inconsistency, Proc. IJCAI 2007, pages 599-604, Jan. 2007, Hyderabad, India.

O. Udrea, D. Reforgiato and V.S. Subrahmanian. Annotated RDF, Proc. 2006 European Semantic Web Conference, Budva, Montenegro, Lecture Notes in Computer Science, Volume 4011, June 2006, Pages 487 - 501, DOI 10.1007/11762256_36, URL http://dx.doi.org/10.1007/11762256_36

J. Dix, W. Faber and V.S. Subrahmanian. The Relationship between Reasoning about Privacy and Default Logics, Proc. 2005 Intl. Conf. on Logic Programming and Artificial Intelligence Reasoning, Dec. 2005, Jamaica.

O. Udrea, Y. Deng and V.S. Subrahmanian. Probabilistic Ontologies and Databases, Proc. 2005 Intl. Conf. on Ontologies, Databases, and Semantics, Oct/Nov 2005, Agia Napa, Cyprus Proc. Springer Lecture Notes in Computer Science, Vol. 3760.

Y. Zhang, E. Manister, S. Kraus and V.S. Subrahmanian. Approximation results for probabilistic survivability, accepted for publication in the Proc. of the 2nd IEEE Symposium on Multi-Agent Security and Survivability, Philadelphia, Aug. 2005. This was the highest rated paper at the conference.

V.S. Subrahmanian, B. Stroe and S. Dasgupta. Optimal Status Sets of Heterogeneous Agents, Proc. 2005 Intl. Conf. on Autonomous Agents and Multiagent Systems, pages 709-715, Utrecht, Netherlands, July 2005. Shortlisted for the best paper award (4 papers out of approx. 530 submissions were shortlisted).

E. Hung, Y. Deng and V.S. Subrahmanian. RDF Aggregate Queries and Views, Proc. 2005 IEEE Intl. Conf. on Data Engineering, Japan, March 2005.

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 12

(d) Manuscripts

Number of Manuscripts: 0.00

Number of Inventions:

Graduate Students

NAME	PERCENT_SUPPORTED
Octavian Udrea	0.75
Edward Hung	0.25
Yu Deng	0.25
Cihan Tas	0.25
FTE Equivalent:	**1.50**
Total Number:	**4**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Venkatramanan Subrahmanian	0.15	No
FTE Equivalent:	**0.15**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME	
Total Number:	

Names of personnel receiving PHDs

NAME	
Yu Deng	
Edward Hung	
Octavian Udrea (to be awarded in summer	
Total Number:	**3**

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Heterogeneous Uncertainty Management

Final Report

ARO Contract DAAD190310202

PI: V.S. Subrahmanian

Dept. of Computer Science,

University of Maryland,

College Park, MD 20742.

vs@cs.umd.edu

Foreword

Uncertainty arises continuously in applications where data needs to be integrated. Mediator technology – followed by agent technology - provides a paradigm within which to integrate data. In this proposal, I studied various problems related to integrating multiple data sources and multiple reasoning paradigms. I started by first looking at *heterogeneous agents* that can be built on top of legacy software code. I developed, jointly with others, the notion of a heterogeneous temporal probabilistic (HTP) agent. HTP agents can build temporal probabilistic reasoning capabilities on top of multiple databases and software packages encoding different reasoning methods. Subsequently, I developed methods to probabilistically reason about how to ensure survivability of multiagent systems. Moreover, at the time, XML provided an important mechanism that could be used to hide the underlying differences between different databases. I proposed, jointly with others, the notion of a of a probabilistic interval XML database in which uncertainty spanning such XML DTDs could be expressed. Later, I extended this to the relatively new web standard, RDF. Along a concurrent path, I also looked at the problem of probabilistic spatio-temporal reasoning and (SPOT) databases. SPOT databases are a paradigm for reasoning about moving objects where there is uncertainty about when and where a vehicle will be. We also proposed the

concept of *pgo*-theories for reasoning about plans of moving objects that are known in advance.

Problem Statement

Reasoning about uncertainty is an important aspect of all real world reasoning today, with specific relevance for the DoD. In DoD applications, uncertainty abounds. We are uncertain about where an enemy vehicle might be – now or in the future. We are uncertain about even the locations of our own vehicles at a given point in time. There is uncertainty in inferences about what these vehicles might be planning to do. This uncertainty also pervades data. There is uncertainty about the quality of data from different sources. There is uncertainty about how to answer queries against such data, and how to effectively deal with answering queries in the presence of spatio and/or temporal and/or data uncertainty. The goal of my work was to develop the theoretical and practical foundations required to integrate multiple forms of data and reasoning when uncertainty was present.

Summary of the most important results

1. **Heterogeneous temporal probabilistic agents.**

We proposed the concept of a heterogeneous temporal probabilistic (HTP) agent. An HTP agent is built on top of a given set of databases or software packages. Each software package is assumed to have a set of API functions whose signature (I/O types) is known. Given an API function *f*, we say that *p:f(a1,…,an)* is a *code call*. Each *ai* can either be a value from the appropriate domain w.r.t. *f*'s signature or a variable over that domain. When all the values are constant, a code call says "Execute function *f* defined in package *p* on the specified values" and return a set (there is no loss of generality in this). Thus, *oracle:sql("Select * from emp where Sal > 50k")* is a code call invoking a database query. A code call atom is an expression of the form *in(X,cc)* where *cc* is a code call. When *X* is a variable, it can be bound to any object returned by the code call *cc*. As an example, the code call atom *in(X, oracle:sql("Select * from emp where Sal > 50k"))* will bind *X* to any tuple returned by the database query expressed above. A *code call condition* is a conjunction of code call atoms and certain atoms called comparison atoms.

For instance, *in(X, oracle:sql("Select * from emp where Sal > 200k")) & in(Y,mapquest:route(X.address,office-address)) & Y.estimatedTime > 30* is a code call condition which tries to find all (X,Y) pairs such that *X* is an employee record about an employee who makes a lot of money (over 250K) but also has a long commute. An action atom is just like an ordinary atom, except that a special action symbol is used in place of a predicate symbol. For example, *drive(X,Y,R)* might be an action atom saying that some agent from drive from X to Y along route R. A *tp-annotated action atom* is an expression that says that the given action will be performed during some time window (specified by a temporal constraint) and that the precise probability that it will be performed at a given time (solution of the constraint) is given by a probability distribution.

An *HTP agent* is a finite set of rules of the from "if some code call condition is true (across the multiple DBs and/or software modules) and if "some tp-atoms are true, then another tp-atom should be true."

We defined the syntax and semantics of HTP agents via the concept of a feasible temporal probabilistic status interpretation. An FTPSI specifies what an agent can do that is in accordance with its rules and in accordance with a change in state that has occurred. We developed algorithms to efficiently represent and manipulate FTPSIs and to compute an FTPSI when the agent needs to determine what to do. We outlined how FTPSIs could be utilized in the case of stock market applications and in electricity markets, though this was not done in sufficient detail to actually use in such applications.

2. **Probabilistic Interval XML.**

XML is a major platform that was proposed about 7-10 years ago as a way of replacing HTML-based web content by some slightly more semantic content. XML markups have subsequently been proposed for a very wide variety of application and data type domains ranging from images to video to music to geospatial and financial information.

We developed some of the first methods to reason about the uncertainty that might be present in XML data. In particular, we proposed the "Probabilistic Interval XML" (PiXML) data model. For instance, in a surveillance application,

we might be able to classify a moving object as a vehicle, but we are not sure if it is a tank (30%) or a truck (70%). If it is a tank, it might be 60% likely to be a T-70 or 40% likely to be a T-72. PiXML deals with the ability to represent and reason about such uncertainty.

We defined the concept of a probabilistic semi-structured schema and a probabilistic instance, and developed two semantics for it – a global possible worlds like semantics, and a more local semantics. We showed that under some assumptions, the two are equivalent. We then extended the relational model of data to include an algebra that manipulates probabilistic XML data.

3. **Probabilistic and temporal probabilistic aggregates.**

In this work, we considered a direct probabilistic database – a database in which entries within tuples have associated probabilities. How can we answer aggregate queries over such databases?

Intuitively, each possible answer to such an aggregate query has a probability. We first developed a declarative semantics to answer aggregate queries to probabilistic databases. Subsequently, we developed a naïve algorithm called GPA to answer aggregate queries over probabilistic databases. This algorithm takes exponential time. We showed that if we have to answer multiple aggregate queries, then we can do much better – rather than processing the aggregate queries sequentially, we can merge commonalities within them and exploit these commonalities in order to gain better performance. Unfortunately, the worst case complexity is still exponential, so we decided to examine the possibility ot using heuristics. We developed several heuristics as well as a prototype experimental implementation where we determined the conditions under which these algorithms worked well.

In a subsequent paper, we extended the above results to the computation of aggregates involving spatio-temporal data.

4. **Probabilistic RDF.**

Over the last few years, the use of RDF as a paradigm for representing knowledge has grown dramatically. RDF and OWL ontologies exist on a wide

variety of topics ranging from genetics to visual sensor data fusion. These are clearly domains that are chock full of uncertainty - image processing programs based on Bayesian analysis often return probabilistic identifications of objects, while relationships between a symptom or disease and the genetic markers a person may also be probabilistic. In order to express such information, we introduced Probabilistic RDF (pRDF) . We defined the concept of a pRDF schema and a pRDF instance . A pRDF-instance extends RDF triples by allowing unconditioned probability distributions over a set of possible values of an RDF triple. We then provided a formal model theoretic semantics for pRDF based on the possible worlds probabilistic logics of Fagin et. al. and showed that we can associate a monotonic function with any pRDF theory --- this function has a least fixpoint that compactly represents the set of all quadruples entailed by the pRDF theory. However, using the fixpoint to answer queries is not always desirable because the size of the fixpoint can be enormous. We developed algorithms to efficiently answer atomic queries with at most one variable. We developed an experimental tested showing that queries can be answered in very small amounts of time (a few seconds) for pRDF instances as large as 100,000 quadruples.

5. **Probabilistic “go” theories.**

There are many applications where one may wish to reason about a set of moving vehicles. For example, DARPA’s CoABS program developed the Coax SYSTE (jointly with the US Navy, Lockheed Martin, BBN, and other companies) ways to predict where and when an enemy submarine would be in the future (and with what probability) based on knowledge about its past movements, terrain conditions, etc. Their predictions consist of a set of statements of the form ``Vehicle V will be at location L with some probability in the interval [L,U]”. Cell phone companies are (and in some cases already have) developed methods to predict where cell phone users will be going in the future --- a small number of law enforcement agencies in the US already use such probabilistic predictions to track selected criminals and such predictions help determine where best to cut them off.

We developed a principled approach to solving such problems by extending ``go'' theories due to Yaman et. al. Their framework is suitable for reasoning

about applications where we know the vehicles' intended destinations --- however, there are many applications such as those mentioned above where this is not known with certainty. A second drawback of the above framework is that while temporal indeterminacy is allowed via intervals, no probability measure is associated with those intervals. This again is appropriate when we are reasoning about plans known to us (e.g. flight plans), but is not appropriate when we are reasoning about a vehicle (e.g. an enemy vehicle on the battlefield) whose plans are not known to us with 100% accuracy.

We proposed ``probabilistic'' go (PGO) theories by building on Yaman's past work. A PGO theory allows us to reason about motion plans that we know as well as motion plans that we do not know with 100% certainty. We developed a syntax and model theoretic semantics for PGO theories. We then showed how to check consistency of PGO theories via linear programming. However, the size of the linear program in question may be exponential, leading one to initially suspect (wrongly) that consistency checking here is NP-complete. We subsequently determine that this problem is polynomially solvable (under the assumption that we are reasoning only about a finite future) by constructing a polynomially sized set of linear constraints for consistency checking and to answer certain kinds of queries called probabilistic "in" queries such as ``is vehicle *id* within a given region at a given time with probability over a threshold?'' Such queries are obviously of great utility. We developed experimentation showing our algorithms perform well in practice.

6. **SPOT databases.**

A SPOT database is a variant of pgo-theories and consists of statements of the form "The probability that a vehicle is within region *r* at time *t* is in the interval [L,U]." In a series of two papers (one published, one undergoing a second round of revisions), we developed several results on SPOT databases.

First, we developed the syntax and semantics of SPOT databases. We developed a naïve algorithm to check consistency of SPOT databases by solving a linear program associated with a SPOT DB. Later, we realized we could do better. We found ways of significantly reducing the size of the linear program being solved, yielding a much faster consistency checking algorithm.

We also studied two types of selection operations. A *cautious select* operator corresponds to the situation where the SPT-atoms in the answer to the select query are somehow entailed by the SPOT DB. In contrast, the optimistic selection operation considers the situation where the SPOT atom is included in the answer if it satisfies the selection condition and is consistent with the answer. We developed algorithms to answer both kinds of queries. We developed a highly specialized index structure called a SPOT-tree by modifying the concept of an R-tree and showed that it would greatly speed up the computation of optimistic selection queries.

We also developed methods to answer join queries across a SPOT DB, and gave a brief outline of how union, intersection, and difference queries could be answered. We also defined an expected nearest neighbor operation that allows us to define which objects are nearest neighbors to which other objects at a specified time with maximal probability.

7. **Probabilistic survivability of multiagent systems.**

One major obstacle to the wider deployment of multiagent systems (MASs) is survivability and reliability. MASs that are deployed across a network can quickly ``go down'' due to external factors such as power failures, network outages, malicious attacks, and other system issues. Protection against such unexpected failures that disable a node is critical if agents are to be used as the backbone for real world applications.

We focused on how replication can form the basis of one tool (amongst many that are needed) to prevent an MAS from succumbing to failure. By replicating agents, we hope to increase the probability that the system will survive the failures, i.e., that at least one copy of each agent will continue to reside on a connected, working host computer, so that the MAS as a whole can function as a unified application.

We built upon past work that defined the probability that a given deployment of an MAS will survive, and addressed two problems. How should we compute the survival probability, given a MAS and its deployment? And how can we find the optimal/most survival deployment, given a MAS, i.e. finding the deployment with the highest survival probability.

We developed an abstract formal model to study the survival probability of a given deployment under the assumption of independence of node failures. We show that the complexity of the most survivable deployment problem, even assuming independence, is at least NP-hard. We also show that for any polynomial approximation to find a sub-optimal deployment, there will be instances in which the survival probability of the most survival deployment is 1 but the algorithm returns a deployment with a survival probability of 0. Thus, any polynomial approximation algorithm is guaranteed to find at least one terrible solution. We introduce two centralized algorithms to accurately compute the probability that a given deployment will survive. Both algorithms take exponential time.

We then developed five different approximation algorithms to compute the probability of survival of a given deployment. We provide a detailed comparison of the performance of the different algorithms proposed in this work. These experiments try to identify the conditions under which one algorithm is preferable to another so that MAS applications have some foundation upon which to base a decision about which algorithm to use.

Technology Transfer

Though none of the work in this proposal was directly transferred to any 3rd parties, some of the work done under this proposal led to follow up work that was transferred.

The concept of an HTP agent led to the subsequent development of a related, but different concept called "Stochastic Opponent Modeling Agents" (SOMA). SOMA was used to build models of several tribes on the Pakistan Afghanistan border. Data on these tribes, collected under separate effort, was shipped to the Army's 10th Mountain Division and was well appreciated. SOMA rules describing the behaviors of several tribes in the same region were shipped to the Army's TRADOC Intelligent Support Activity and to Army/AMSAA. In both cases, the results were well received.

Bibliography

Y. Zhang, E. Manister, S. Kraus, V.S. Subrahmanian, D. Peleg. Computing the survivability of a multi-agent deployment, Artificial Intelligence journal, accepted subject to revisions, Oct. 2007.

A. Parker, J. Grant, and V.S. Subrahmanian. A Logical Formulation of Probabilistic Spatial Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, Nr. 11, pages 1541-1556.

M. Fayzullin, V.S. Subrahmanian, M. Albanese, C. Cesarano, A. Picariello. Story Creation from Heterogeneous Data Sources, Multimedia Tools and Applications, Vol. 33, Nr. 3, pages 351-377, June 2007.

E. Hung, L. Getoor and V.S. Subrahmanian. Probabilistic Interval XML, ACM Transactions on Computational Logic, Vol. 8, Nr. 4, August 2007.

J. Dix, S. Kraus and V.S. Subrahmanian. Heterogeneous Temporal Probabilistic Agents, ACM Transactions on Computational Logic, Vol. 7, Nr. 1, Jan. 2006.

R. Ross, V.S. Subrahmanian, and J. Grant. Aggregate Operators in Probabilistic Databases, Journal of the ACM, Vol. 52, Nr. 1, pages 54-101, January 2005.

J. Grant, G. Infantes, A. Parker and V.S. Subrahmanian. SPOT Databases: Efficient Consistency Checking and Optimistic Selection in Probabilistic Spatial Databases, submitted to a technical journal, Nov. 2007. Under revision in accordance with reviewers' comments.